学校编码：10384　　分类号_____密级_____
学号：X2011231120　　UDC _____

厦门大学

工 程 硕 士 学 位 论 文

基于数据挖掘的中小地税机构税收预测系统分析与设计

Analysis and Design of Revenue Forecasts Organization System for Small and Medium-Sized Local Taxation Bureau Based on DM

沈立鸿

指 导 教 师：　王美红 助理教授
专 业 名 称：　软　件　工　程
论文提交日期：　2013 年 10 月
论文答辩日期：　2013 年 11 月
学位授予日期：　　年　月

指 导 教 师：_____
答辩委员会主席：_____

2013 年 10 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

随着信息化技术的发展，各级税务部门可获得的税收征管数据容量越来越大，海量的税收征管数据背后隐藏许多重要信息。

在传统的税收预测工作中，中小地税机构无法有效的集成各个“信息孤岛”，同时缺乏发现税收征管数据背后隐藏知识的手段，从而出现“数据爆炸而知识贫乏”现象。税收预测人员希望对海量的税收征管数据进行更深层次的分析，生成与税收预测相关性更大的税收征管数据，为税收预测和决策提供帮助。

本论文就 Visual Basic for Applications (VBA，Visual Basic 的一种宏语言)结合数据挖掘技术在税收预测中的应用问题，在掌握和分析关于数据挖掘技术发展、传统税收预测技术的基础上，对税收征管数据挖掘过程模型建立、数据预处理技术、关联规则挖掘等问题进行研究，主要完成工作如下：

1. 中小地税机构税收预测系统的需求分析，结合税收征管数据处理流程和业务需求，提出了中小地税机构税收预测系统的设计框架；

2. 在了解税收征管数据存储和处理现状的基础上，针对税收征管数据中存在的问题和数据挖掘技术对数据的要求，总结了常见的数据预处理技术和适用情况；

3. 数据挖掘组件的设计，针对税收征管数据特点调用 SQL Severe 2000 中的 Microsoft 决策树和 Microsoft 聚集方法进行数据挖掘；

4. 税收预测组件的设计，针对税收预测人员习惯，采用 VBA 方式在在 EXCEL 2003 中实现基于时间数列、基于税收与经济关系的预测方法；

5. 其它相关组件的设计。

该系统主要功能经税收预测人员试用，使用效果良好。这些预测结果可以帮助税收预测人员做进一步预测，从而提高税收预测的准确性，使低层次的简单查询、统计税收征管数据应用提升到知识挖掘和决策支持的高级应用。

关键词：数据挖掘；税收预测；中小地税机构

Abstract

With the development of information technology, the tax authorities at all levels can obtain the tax collection and administration of data capacity is more and more big, huge amounts of data of tax collection and administration, behind a lot of important information.

In the traditional tax forecasting work, integrate the "information island" and tax agencies cannot effectively, while the lack of tax collection and management of data hiding behind the discovery of knowledge means, thus appeared "data explosion but knowledge poor" phenomenon. Tax forecasting researchers hope to tax collection and management of data on mass for a deeper analysis, correlation between generation and revenue forecast more tax collection data, to provide help for the tax revenue forecasting and decision.

Application of the Visual Basic for Applications (VBA) combined with data mining technology in the revenue forecast, in the control and Analysis on the basis of the development of data mining technology, traditional tax forecasting technology, process modeling, data preprocessing, association rules mining, mining to the tax revenue collection data for research, the main work completed is as follows:

1. small and medium-sized local tax bureau tax forecasting system requirements analysis, combined with the tax collection and management of data processing and the business demand, proposed the small local tax bureau tax forecasting system design framework;

2. On the basis of tax collection and management of data storage and processing status of understanding, according to the requirements and the data mining technology in tax collection and management of data in the data, the data preprocessing technology common and applicable condition;

3. Component design, data mining, according to the tax collection and management characteristics of data called SQL Severe 2000 Microsoft decision tree and Microsoft aggregation methods for data mining;

4. Component design for prediction of tax, tax forecasters habits, using VBA mode in EXCEL 2003 based on time series prediction method, based on tax and economic relationship;

5. Design and other related components

The main functions of the system had reviewed by tax forecasting personnel,which obtained a good evaluation. These results can help researchers to do further forecast revenue forecast, so as to improve the accuracy of forecast of revenue, the low level of simple query, statistics application of tax collection and management data to knowledge mining and decision support advanced applications.

Key Words: Data Mining；Tax Forecasting；Small and Medium-Sized Local Taxation Bureau

目 录

Contents

第一章　绪论

1.1　引言

随着信息化技术的发展，极大方便了人们对信息的获取和使用，但同时也让不少人在面对海量信息的时候显得手足无措。人们想要快速、准确地在这些信息中找到所需要的信息，发现数据隐含的、未知的、有价值的潜在信息变得越来越困难。于是，数据挖掘（DM）被提了出来，为决策人员提供辅助。

数据挖掘在各领域的应用非常广泛，只要该产业拥有具分析价值与需求的数据仓储或数据库，皆可利用数据挖掘工具进行有目的的挖掘分析，目前多应用在零售业、直销界、制造业、财务金融保险、通讯业以及医疗服务等。常见的成功案例有数据挖掘帮助 Credilogros Cía Financiera S.A.改善客户信用评分、数据挖掘帮助 DHL 实时跟踪货箱温度、“尿布与啤酒”等。

税收工作信息化的标志性事件，是美国国内收入局（Internal Revenue Service,IRS）在 1969 年引入的选案系统“判别清单函数系统”（Discriminate Inventory Function System,DIF），美国的税务征收管理工作（在美国称为税务审计）从此开始走上主要依靠计算机选案的现代化之路。DIF 系统是一种数学技术，它是通过寻找给所得税申报表打分，从而找出具有检查潜力的申报表来审计。DIF 系统之所以能够一直被人青睐，和该系统的数据库质量高、数据更新及时、数据覆盖面广、数据来源广是密不可分的。

为了跟上时代发展的步伐，我国税务部门的信息化建设水平也在不断提高，而面对不同时间上线、不同开发商开发、不同业务部门使用的各种税务管理信息系统产生的大量数据，如何从中发现隐含的知识成为了一个难题。近几年来，税务总局与部分发达地区已经对如何运用数据挖掘做出了相应的尝试与努力。但是，因为它们所处的层级与地区经济差异，使得它们的成果并不能简单的套用到基层中来，更多的时候只能作为一个指导性意见而存在。

1.2 课题的目的和意义

税收是国家为实现其职能，凭借政治权力，按照法律规定，通过税收工具强制地、无偿地征收参与国民收入和社会产品的分配和再分配取得财政收入的一种形式。2012 年全国公共财政收入 117210 亿元，税收收入完成 110740 亿元（不包括关税和船舶吨税，未扣减出口退税）、占国家财政收入的 86%，比上年同期增长 23.43%。能够实现如此大的增幅，税收预测功不可没。

税收预测是以充分掌握影响税收收入变动的因素和税收历史资料为基础，以统计方法、数学方法为手段，经过严密的推理和计算，对未来税收收入的前景做出比较确定的判断的一项科学管理工作，其目的在于分析和预见税源、税收和税务工作未来的发展变化趋势。税收预测工作对于加强组织收入工作，更好地完成税收任务，为领导科学决策和管理提供服务等，都具有重大意义[1]。它既是税务部门制订税收计划的科学依据，也是国家及各级政府预算的重要参考。

税收信息化起源于上世纪 80 年代后期，从单机税收会计运用开始。到上世纪 90 年代初，税收征管系统初步建立起来，它们主要是将手工操作流程转化为自动处理流程。从上世纪 90 年代末开始，国内的信息技术开始了质的提高，税收信息系统开始经历从区县集中发、地市集中、省局集中再到全国集中的历程。随着税务部门信息化建设水平的不断提高，不同时间上线、不同开发商开发、不同业务部门使用的各种税务管理信息系统产生了大量的数据。然而，这些数据在中小地税机构中，更多的时候却成了一些“信息孤岛”，缺乏有效的集成，它们很难为管理层的决策支持作出比查询更多的贡献。

特别是在这些机构中，原本就普遍存在着人员偏少、人员结构不合理等的情况下，这些“杂乱”的数据就更难以被有效利用起来。同时，这些中小地税机构还必须面对着，宏观上的预测数据并不能简单、直接的应用到税收收入有着巨大地区差距的基层部门中来。

而数据挖掘刚好是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取正确的、有用的、未知的、综合的以及用户感兴趣的知识并用于决策的过程，税务总局与部分发达地区已经对于如何运用数据挖掘做出了相应的尝试与努力。但是，因为它们所处的层级与地区经济差异，使得它们的成果并不能简单的套用到基层中来，更多的时候只能做为一个指导性意见而存在。

如何转化上级部门的成果为己用，如何利用好自身积累的数据，都成为了信息化建设中必须面对的问题。

1.3 研究现状分析

1.3.1 数据挖掘研究现状

数据挖掘（Data Mining），又称为数据库中的知识发现（Knowledge Discovery in Database,KDD）。KDD 一词首次出现在 1989 年 8 月举行的第 11 届国际联合人工智能学术会议上。

1993 年，IEEE 的《Knowledge and Data Engineering》会刊出版的 KDD 技术专刊所发表的 5 篇论文，较全面地论述了 KDD 系统方法论、发现结果的评价、KDD 系统设计的逻辑方法，集中讨论了鉴于数据库的动态性冗余、高噪声和不确定性、空值等问题，KDD 系统与其它传统的机器学习、专家系统、人工神经网络、数理统计系统的联系和区别，以及相应的基本对策[2]。它们成功的展示了 KDD 从建立模型到设计实现的具体应用方法。

如今，美国人工智能协会主办的 KDD 国际研讨会规模已经由原来的专题讨论会发展到国际学术大会，研究重点也从发现方法转向系统应用，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。此外，数据库、人工智能、信息处理、知识工程等领域的国际学术刊物也纷纷开辟了 KDD 专题或专刊，在 Internet 上还有不少电子出版物。

尽管数据挖掘作为一个独立的学科才有十多年的时间，但随着计算机硬件与软件的发展，它也被应用到越来越多的领域。目前，国外数据挖掘的最新发展主要有对发现知识的方法的进一步研究，如近年来注重对贝叶斯（Bayes）方法以及 Boosing 方法的研究和改进提高、KDD 与数据库的紧密结合、传统的统计学回归方法在 KDD 中的应用。在应用方面主要体现在 KDD 商业软件工具从解决问题的孤立过程转向建立解决问题的整体系统[3]。Oracle、Microsoft 和 IBM 等主流的数据库厂商，已在其产品中增加了数据挖掘功能。

与国外相比，国内起步稍晚。数据挖掘在上世纪 90 年代进入了中国，1993 年国家自然科学基金首次支持对该领域的研究项目。随后，国内的许多科研单位、

高等院校都参与到相关研究工作中来，并在模糊方法的知识发现、关联规则挖掘算法、WEB 数据挖掘等方面取得了一定的成果。

目前，国内的数据挖掘技术应用主要集中在电信的客户分析、农业(行业数据预测)、零售(销售预测)、网络日志(网页定制)、电力、银行(客户欺诈)、生物(基因)、天体(星体分类)、医药、化工等方面[4]。它能解决的问题主要在于：数据库营销(Database Marketing)、客户群体划分(Customer Segmentation & Classification)、背景分析(Profile Analysis)、交叉销售(Cross-selling)等市场分析行为，以及客户流失性分析(Churn Analysis)、客户信用记分(Credit Scoring)、欺诈发现(Fraud Detection)等等，并且在许多领域都得到了成功的应用。

1.3.2 数据挖掘技术在税务系统的应用现状

早在 1998 年，美国加州税务启动的基于 IBM DB2 数据库软件的综合逃税人监察项目数据仓库解决方案（INC）项目，使加州税务能够在超过 2.2 亿项的独立税务信息中利用商业智能技术进行业务分析。

国内自从 1994 年税制改革以来，为适应经济环境的不断变化，税收法律和制度建设的步伐不断加快。在探索中前进的税收征管软件不得不时时调整，从而带来历史征管数据垃圾较多、缺乏数据仓库的具体应用规划的问题。

对此国家税务总局已经对原有问题进行了研究，并通过金税工程提出通过全国统一标准的信息化平台，实现全国税收业务统一、税务行政管理规范、强化税收执法权和管理权的监督制约，推动税收事业全面发展。目前，金税工程已经进行到第三期工程，初步建立了全国集中管理的统一性、规范性的涉税数据库，涉税信息在各部门、各涉税环节可以顺畅流转、多次复用，并能够实现各数据间的交叉审核和流程监控；将建成能够满足税务部门各级、各层次管理者多样性目标的、灵活高效的数据分析和应用平台，实现各级管理人员对全国税务信息的实时查询和监控、分析，为决策分析提供完整、准确、及时的数据源，为各级税务机关税收决策提供依据。